

Фактограф: система для создания, поддержания и публикации электронных архивов*

А.Г. Марчук

Институт систем информатики им. А.П.Ершова СО РАН
mag@iis.nsk.su

Аннотация

В работе представлен подход к формированию цифровых архивов документов. Он отличается тем, что кроме основных единиц хранения электронных образов документов, создается база данных неспецифических сущностей, к которой «привязываются» документы. Такой способ позволяет улучшить описание документов и привести естественные формы поиска и изучения материала. В работах [1, 3] определены принципы электронной фиксации фактов, а также общие положения реализации распределенных архивных фактографических систем. Усилиями последних лет было создано платформенное решение Фактограф, позволяющее создавать, сопровождать и публиковать электронные архивы. Систему также можно использовать и в других прикладных областях: цифровые библиотеки, музеи, простые системы научного делопроизводства.

1. Введение

Созданием электронных архивов, коллектив Института систем информатики им. А.П.Ершова СО РАН занимается около 15 лет. Работы носят фундаментальный характер по исследованию принципов и способов создания таких архивов, но также за все эти годы выполнялись конкретные разработки прикладных архивов. Задача состоит в том, чтобы множество документов той или иной природы упорядочить в систему, способствующую удобству доступа к документам и сопровождающей документы дополнительной информации. Под документами понимаются и традиционные хранимые рукописные, машинописные материалы, так и фотографии, видео, аудио. В оцифрованном виде сюда подходят и музейные экспонаты, электронные публикации в разных форматах и т.д.

На первый взгляд, работа могла бы свестись к сканированию и другой обработке элементов архива с выкладыванием полученной информации в Ин-

тернет с упорядочением по стандартному для архивного дела принципу временной зависимости документов. Однако даже небольшие архивы содержат довольно большое количество документов и поиск нужного в таком случае становится затратным действием.

С самого начала прикладные разработки велись в областях, где отсутствует устоявшаяся система классификации, поэтому была взята ориентация на создание сопряженной базы данных, связывающей документы с элементами базы данных, такими как темы, коллекции, информация о персонах, событиях и географических точках. В случае переработки "бумажного" архива в электронную форму, создание базы данных становится весьма трудозатратной. Например, формирование базы данных Электронного архива академика А.П.Ершова было закончено лишь через 7 лет после начала проекта, тогда как сканирование материалов (более 500 толстых канцелярских папок) и создание информационной системы, было выполнено достаточно быстро. Тем не менее, было выявлено, что такой подход к созданию электронных архивов дает пользователям-исследователям новое качество работы с документами архива. Это не только быстрая доступа к отдельным документам, но и возможности получения ссылок на группы документов по различным признакам, напр. по связи документов с тем или иным лицом. Кроме того, выявилось, что созданная база данных может использоваться и самостоятельно, например, база данных участников конференций, данные о произошедших событиях, о персонах и организациях. В работе над другим большим проектом — фотоархивом СО РАН, база данных носит принципиальный характер, поскольку основные документы это фотографии, интерпретация которых важна для понимания.

Поскольку документы в архивном плане важны не как красивые картинки, а как носители информации, раскрытие смысла является ключевым направлением в проводимых исследованиях. Это выразилось и в том, что уже около 10 лет в работе используются концепции и стандарты Semantic Web [2]. В частности, база данных накапливается в формате RDF, создана и используется онтология. К сожалению, в исследованиях и проектах не удалось воспользоваться внешними разработками по ряду причин. Было сделано несколько вариантов своего программного обеспечения, которое в совокупности было названо "Фактограф", но по существу, пред-

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

ставляет собой набор средств или платформу для создания архивных систем. Смена программных решений не оказала на ведущиеся проекты болезненного влияния, поскольку основные фактографические принципы, онтология и база данных практически не менялись, что позволяет накапливать данные не только в рамках одного проекта, но и интегрировать данные и накапливать уже более широкий набор информации. В основном, наши проекты идут по материалам деятельности Сибирского отделения РАН и его людей. Так что, в совокупности накапливается информация об истории и социуме Сибирского отделения.

2. Общие принципы

Предлагаемая информационная система допускает следующие основные виды работ: ввод и первичное упорядочивание отсканированных документов и других документов в электронной форме, публикацию документов в Интернет в соответствии с ограничениями Интернета и политикой владельца документа, ввод и редактирование базы данных различных сущностей, аннотирование документов, формирование универсальных или специализированных интерфейсов к создаваемому архиву, использование электронного архива в виде сервиса со стороны других информационных систем и сайтов. Кроме того, система допускает распределенный характер накопления информации по сохранении интеграции отдельных архивов. Под интеграцией архивов понимается то, что одни и те же сущности (люди, организации, события) были бы в архивах в едином варианте и чтобы та информация, которую авторы архива считают общезначимой, была бы доступна и другим участникам объединения архивов и подвергалась совместному пополнению и редактированию.

Таким образом, Фактограф позволяет не только выполнять локально работу по созданию "своего" архива, но и позволяет выполнять коллективную работу с вовлечением в нее разных групп, обладающих разными документными массивами. В мире создаваемых текстовых коллекций, для обеспечения совместной работы над общим информационным полем, в основном используются технологиями Wiki. В нашем случае, эти технологии не применимы, сформирован и создан свой подход к совместному накоплению данных. Суть его была сформулирована методологами Semantic Web как "любой о любом может сказать любое". В принципе, как в Википедии, но применительно к строго структурированной базе данных. Главной особенностью нашего решения совместного редактирования, является проецирование деятельности информационного оператора на "его" часть базы данных при том, что ему позволено редактировать и "чужие" данные. В случае некомпетентности или злого умысла со стороны редактирующего субъекта, его накопленные данные можно просто отключить.

Использование подходов Semantic Web позволяет разные части "осмысленной" информации создавать в разных системах понятий. С одной стороны, это решает проблему специфичности той или иной фиксируемой предметной области, с другой стороны, это же препятствует интеграции данных и "взаимопониманию" информационных систем, работающих на общем информационном поле. В фактографе используется промежуточный подход, заключающийся в том, что для архивных систем рекомендуется базовая онтология BONE, как общая система понятий и структуризации. При этом, позволено расширить онтологию в сторону "своей" предметной области и формировать архив в расширенной онтологии. Надо только понимать, что специфичная часть информации будет хоть и доступна в рамках обобществления информации, но не всегда будет способствовать интеграции архивов и не всегда будет "понята" другими информационными системами.

3. Архитектура системы

Система базируется на ряде соглашений и формальных конструкций. Они выработались в нашем коллективе достаточно давно, в целом удовлетворяют потребности по решению конкретных задач и по эволюции подходов. Основой применяемых решений являются стандарты и рекомендации группы Semantic Web [4]. Ряд особенностей подхода описан в [5, 6].

3.1. RDF-файлы

База данных формируется как совокупность RDF-файлов, создаваемых, редактируемых и используемых в рамках накопления информации. Используется стандартный вариант представления RDF в виде XML-документа. Для того, чтобы подчеркнуть специфичность формата, такие RDF-документы в системе называются FOG-документами. Присутствуют определенные ограничения, такие как отказ от неименованных сущностей (Blank Nodes), встроенных механизмов bag, alt, list, reification, используется только одна схема идентификации узлов. В принципе, это не привносит ограничений, хотя сочетать внешние RDF-документы с FOG-документами пока мы не пробовали.

Важным и принципиальным ограничением, присутствующим в FOG-структуризации является то, что из одного узла, за исключением оговоренных ситуаций, не допускается более одной дуги (объектной связи или поля данных) с одноименными идентификаторами. По полям данных исключением являются языковые варианты поля, напр. имя по-русски, имя по-английски и т.д. В принципе, такое ограничение приближает RDF-структуризацию к классической системе реляционных таблиц, причем в наиболее ее эффективном варианте, когда есть первичные ключи (идентификаторы сущностей) и внешние ключи (объектные ссылки).

Одновременно, появляется возможность решить задачу, на которые мало кто из исследователей обращает внимание - это редактирование "чужих" данных. В понятие редактирования входит не только добавление сущностей и связей, но и их изменение и уничтожение. Стандартный RDF не дает возможностей изменять и уничтожать "чужие" определения. В нашем случае, все исходящие от сущности свойства (properties) можно сгруппировать в запись. Причем к записи добавляется отметка времени создания. И тогда внесение нового варианта записи в "свой" FOG-документ, подменяет предыдущее определение. А уничтожение сущностей осуществляется с помощью явного оператора delete с указанием идентификатора уничтожаемой сущности. Кроме этого расширения, введен также оператор substitute, в котором явно указывается какой идентификатор сущности заменяется на какой. Этот оператор соответствует отношению sameAs, имеющемуся в OWL, но прямо указывает на то, какой идентификатор теперь становится оригиналом.

3.2. Онтология

Центральным звеном в системе соглашений Фактографа является онтология. Система построена так, что онтология быть сменной. Однако имеется группа особенностей предметной области, которые предписывают фиксировать схему, по которой структурируются данные. Такая онтология была разработана для задач общего назначения. Это онтология BONE (Basic Ontology for Non-specific Entities).

Определяемыми классами онтологии являются персоны, организационные системы, документы, географические системы и коллекции. Между ними определены различные отношения такие как участие, отражение, элемент коллекции, расположение, авторство, титулы, а также естественные атрибуты (DatatypeProperties) имя, начальная дата, конечная дата роль, комментарий и др.

На рисунке 1 в древовидной форме представлен фрагмент онтологии. Изображена иерархия основных

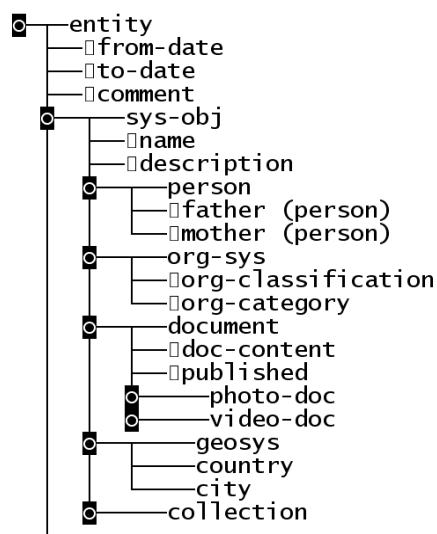


Рис. 1. Основные сущностные классы

сущностных классов вместе с используемыми полями (атрибутами) и ссылками, помещенными прямоугольниками. В случае ссылки (ObjectProperty), в скобках указывается класс целевого объекта.

Существенным, и отличным от многих известных онтологических наборов, является подход к реализации средствами RDF отношений. Как правило, отношения формируются в виде идентифицированных псевдосущностей. Это необходимо для атрибутирования отношений. Например, отношение «работа» между персоной и организацией может иметь такие атрибуты как с какого и по какое время, должность, классификация должности. Другой фрагмент онтологии BONE, определяющий такие сложные отношения, представлен на рисунке 2.

3.3. Кассеты

Первичной задачей электронного архива является хранение документов, т.е. файлов, представляющих эти документы. Известные решения - файловые хранилища. Они достаточно хорошо проработаны и выполняются в виде одноуровневых или двухуровневых репозиториях. Такие схемы, при всей своей надежности, недостаточно гибкие, довольно громоздки в использовании, а часто и дорогостоящие. Все это простимулировало к созданию своего простого хранилища документов. Эту конструкцию назвали кассетами. Кассеты реализуются средствами стандартной операционной системы в виде директории с поддиректориями и с регламентированной схемой размещения помещаемых файлов. Помещаемые в кассету файлы документов регистрируются в этой же кассете в простой базе данных. Роль носителя базы данных для кассеты, также как и для всего проекта выполняет FOG-файл, помещенный в специальное место. Другие FOG-файлы базы данных помещаются как обычные документы.

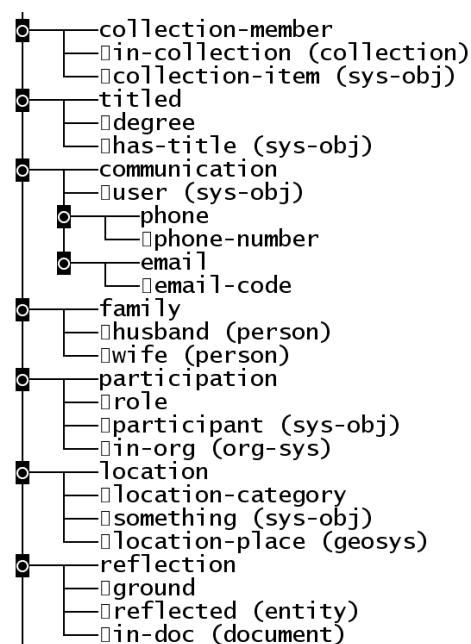


Рис. 2. Основные отношения онтологии

В кассете также реализуется подготовка документов к визуализации через Интернет. Кроме сохраненных оригиналов, большие фотографии преобразуются дополнительно в несколько вариантов типоразмера, видео файлы преобразуются в варианты потокового видео. Подобная подготовка осуществляется и с некоторыми другими видами файлов и наборов файлов. При помещении файлов в архив, полностью сохраняется оригинал файла, за исключением имени файла, формируется служебное имя, под которым файл фиксируется в кассете. Дело в том, что пользователи часто используют имя файла для того, чтобы поместить в него дополнительную информацию, часто на родном языке. Это оригинальное имя сохраняется в метайнформационном поле зафиксированного документа. Кроме служебного имени, размещенный в кассете документ приобретает и логическое имя, которое по желанию пользователя может быть назначено специальным архивным или именем исходного файла.

Как уже указывалось, кассеты это множество файлов и директорий, начинающихся от одного корня, размещенных в файловой системе компьютера. В принципе, это является удобным для администрирования, особенно малоподготовленными пользователями. Средствами операционной системы и внешними утилитами можно осуществлять перемещение кассет, резервное копирование, синхронизацию кассет, размещение и использование в Intranet и др. Но это же может оказаться проблемой если пользователь начнет самостоятельно менять файлы и структуру файлов внутри кассеты. Так что, в будущем предполагается упаковать кассету в какой-то простой конверт. Тогда для пользователя кассета будет представляться одним файлом. Но в силу

сложившейся ситуации в области файловых систем, это не всегда возможно, поскольку размер кассеты в проектах часто превышает, иногда существенно, размер в 4 Гб.

3.4. Эволюция данных и технологий

Архивная система предназначена для того чтобы сохранять документы и данные в течение десятилетий. При нынешних темпах изменений в информационных технологиях, это представляет обширное множество проблем. Мы исходим из того, что среда исполнения будет меняться, программы обработки будут меняться, форматы документов также будут меняться. Архивные данные и документы должны сохраниться и быть доступными для использования. Базовым решением, позволяющим надеяться, что это удастся сделать, является опора на стандарты, которые проживут еще много десятилетий, а также то, что хранимые структуры имеют простое формальное описание.

4. Прикладные проекты

Многие описанные решения были использованы при выполнении ряда прикладных проектов, таких как: Электронный архив академика А.П.Ершова, Информационная система поддержки кафедры программирования ММФ НГУ, Фотоархив СО РАН, Архив Школы юных программистов, Юбилейный портал ММФ НГУ и др. В настоящее время выполняется группа проектов под общей темой «Открытый архив СО РАН». Используя описываемую платформенную систему, создан инструментарий для формирования и редактирования архива, один из Web-интерфейсов изображен на рисунке 3.

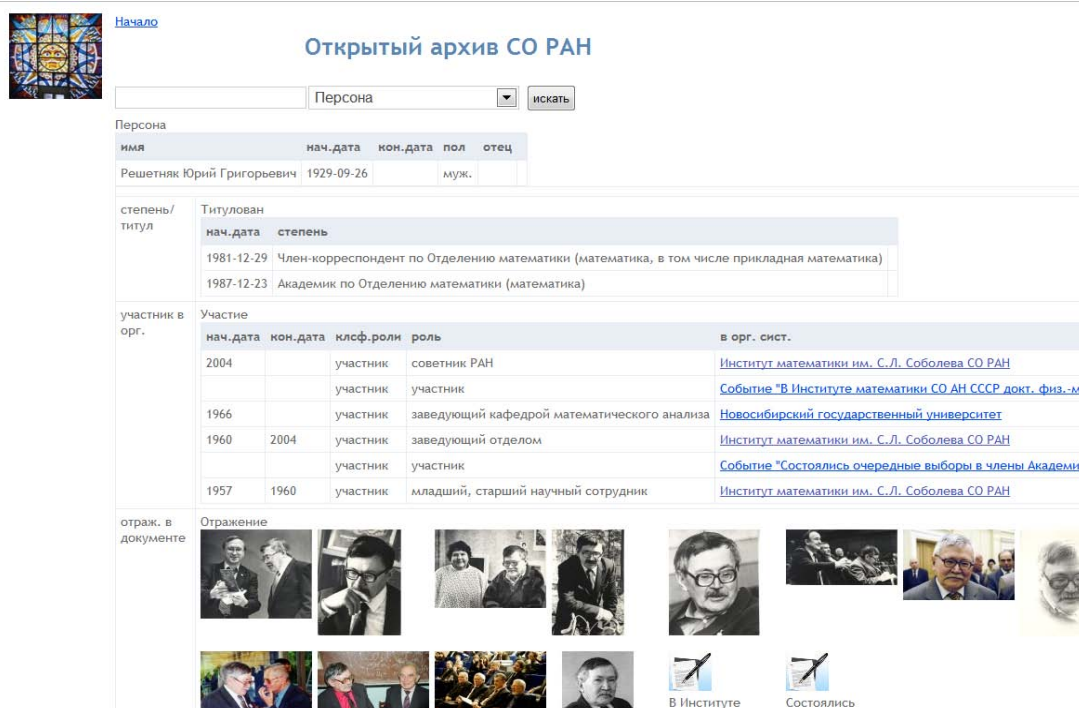


Рис. 2. Интерфейс просмотра и редактирования архивной системы

Литература

- [1] Марчук А.Г. Распределенные электронные архивы, библиотеки и базы данных. Препринт 122 / Институт систем информатики им. А.П. Ершова СО РАН. Новосибирск, 2004. 25 с.
- [2] Berners-Lee Tim, Hendler James, Lassila Ora. The Semantic Web // Scientific American. Volume 284(5). 2001. P. 34-43.
- [3] Марчук А.Г. Методы и технологии исторической электронной фактографии / доклад на конференции ЗОНТ, 2007, публикация в Lecture Notes in Artificial Intelligence.
- [4] Semantic Web. URL: <http://www.w3.org/standards/semanticweb>.
- [5] Марчук А.Г., Марчук П.А. Платформа интеграции электронных архивов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды девятой всероссийской конференции. Переславль, 2007. С. 89-94.
- [6] Марчук А.Г. и др. Модели и методы построения информационных систем, основанных на формальных, логических и лингвистических подходах / Рос. акад. наук, Сиб. отд-ние, Ин-т систем информатики им. А.П. Ершова. Новосибирск: Изд-во СО РАН, 2009. 330 с.

Factograph: System for Creation, Support and Publishing of Digital Archives

A.G. Marchuk

Digital Archives deal with digital copies of documents, photo, video, audio etc. Traditional form of digital document's structuring is database based on meta-data fields like Dublin Core. A new approach for digital archiving is presented in this paper. This approach consists from two parts: structuring not only documents, but also joint important information and second – following ideas and standards of Semantic Web to provide integration and interoperability of local domains of archive data. A platform solution was designed to support the approach. Several applied systems oriented on accumulation of historical documents and facts were build.

* Работа поддержана грантом РФФИ 11-07-00388а, программой РАН Р-15/10, интеграционным проектом СО РАН М-48.